HOW TREHALOSE PROTECTS DNA IN THE DRY STATE:

A MOLECULAR DYNAMICS SIMULATION

A Thesis

by

XUEBING FU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2008

Major Subject: Chemistry

HOW TREHALOSE PROTECTS DNA IN THE DRY STATE:

A MOLECULAR DYNAMICS SIMULATION

A Thesis

by

XUEBING FU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Yi Qin Gao
Committee Members,	Pingwei Li
	Simon W North
	Robert R Lucchese
Head of Department,	David H Russell

August 2008

Major Subject: Chemistry

ABSTRACT

How Trehalose Protects DNA in the Dry State:

a Molecular Dynamics Simulation. (August 2008)

Xuebing Fu,

B.S., Shandong University

Chair of Advisory Committee: Dr. Yi Qin Gao

Molecular dynamics simulations were conducted on a system consisting of a decamer DNA solvated by trehalose and water (molecular ratio= 1:2), to mimic a relatively dry state for the DNA molecule. Simulations were performed at two different temperatures, 300 K and 450 K. The B-form DNA structure was shown to be stable at both temperatures. The analysis of hydrogen bonds between trehalose/water and DNA revealed that trehalose and backbone DNA formed the largest number of hydrogen bonds and thus constituted the major effect of structural protection for DNA. The number of hydrogen bonds formed by each OH group of trehalose with the backbone DNA was compared. Different types of trehalose-DNA interactions were analyzed, with no prevalent pattern recognized. Diffusion constants for trehalose and water were also calculated, suggesting a glassy/viscose state of the simulation system. It is believed that trehalose protects DNA in the dry state through the network of hydrogen bonds built by the sugars, which reduces the structural fluctuations of DNA and prevents its denaturation.

TABLE OF CONTENTS

LIST OF FIGURES

Page

I. INTRODUCTION

A. Overview of molecular dynamics method

Since World War II, there have been magnificent advancements in computer technology, which brings the prosperity of computer simulations at all different levels, ranging from quantum level based on the motion of electrons, to coarse-grained level looking at units as large as macromolecules.

The method we used in this paper is called molecular dynamics (MD), which simulates molecular systems at an atomic level. The MD method has been widely used to investigate the structure and dynamics of biomolecular systems, e.g. proteins, nucleic acids, and small molecules like amino acids and sugars. This method is based on Newtonian dynamics: the potential energy is written as a function of atomic coordinates, and the forces on individual atoms are directly related to the gradient of the energy function; therefore these potential energy functions are commonly referred to as "force fields" [1]. The force fields are a set of simple classical functions, with a large number of parameters adjusted to optimize agreement with experimental data and with quantum calculations on smaller molecules.

The most commonly used force field is written as equation 1:

$$V(r) = \sum_{bonds} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{tortions} k_\phi [\cos(n\phi + \delta) + 1] + \sum_{\substack{nonbond \\ pairs}} [\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}] \quad (1)$$

The first three terms in the summation correspond to force fields over bonds (1-2 interactions), angles (1-3 interactions), and torsions (1-4 interactions). These are shown in

This thesis follows the style of *Journal of Chemical Physics*.

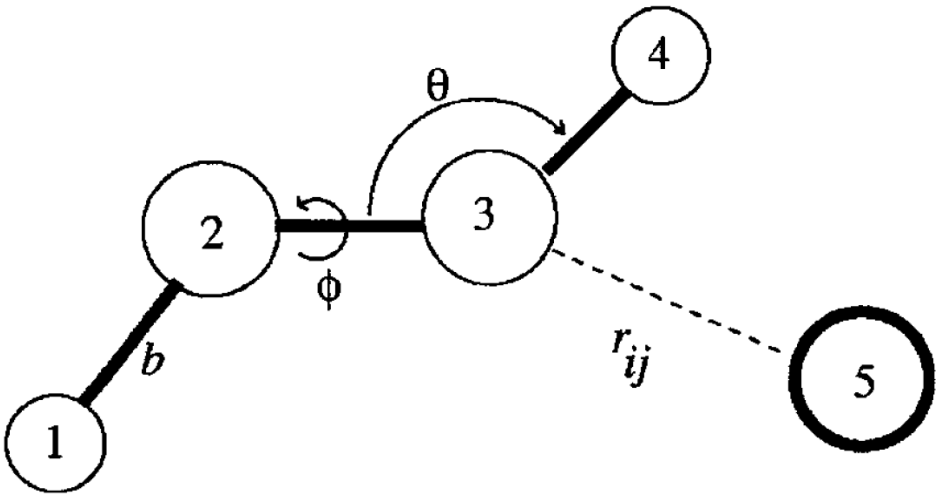

Figure 1. Schematic view of force field interactions. (Covalent bonds are indicated by heavy solid lines, and nonbonded interactions by a dashed line [1].)

Figure 1. The torsion term can also include so-called "improper" torsion terms, which primarily serve to enforce planarity around sp^2 central atoms; other forms of function can be used for the same reason. The last term of the summation (over pairs of atoms i and j) refers to the nonbonding interactions, which represents the electrostatic energy, defined by Coulomb's law, and the "van der Waals" energy, written in the form of a Lennard-Jones 6-12 potential.

With the force field functions, the potential energy of a macromolecular system can be determined as long as its configuration is known. The forces applied on each individual atom can be determined by

$$f_i = -\frac{\partial V(r)}{\partial r_i} \tag{2}$$

According to Newton's law, we have

$$v_i = \frac{dr_i}{dt} \tag{3}$$

and

$$\frac{dv_i}{dt} = \frac{f_i}{m_i} \tag{4}$$

Equation 3 and 4 are ordinary first-order differential equations and can be solved simultaneously by several integration methods. One of the most efficient algorithms for solving the two equations is known as Verlet algorithm [2]. This algorithm leads to the solution of a new position and a new velocity after a very short time step Δt:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{\dot{f}(t)}{2m}\Delta t^2 \tag{5}$$

$$v(t+\Delta t) = v(t) + \frac{\dot{f}(t) + \dot{f}(t+\Delta t)}{2m}\Delta t \tag{6}$$

Since this method stores the position, velocity, and force all at the same time, it minimizes the round-off error and provides numerical stability, convenience and simplicity [2].

Although the MD simulation method is very powerful, in many cases when the energy barrier of a certain process we are interested in is high, it would be very hard/slow to capture this process by running a normal MD simulation introduced above. The main problem is that the phase space points sampled by the normal MD method would mainly focus on those points with very low energy; for the system to overcome the high barrier, it requires high energy. Therefore, the low energy region of the space is overly sampled while the high energy sampling is very scarce. To enhance the sampling of high energy regions, high temperature simulations can be used. However, simulations at higher temperature would be too much focused on the high energy part in the energy space, while the low energy/normal condition part, which is equally important, can be inadequate. So people came up with the idea that if we combine low temperature and high temperature samplings together, then we can get the overall information we want. This is the basic idea of a large series of enhanced sampling methods, e. g. replica exchange method [3], which directly varies the temperature ranges of simulation, and e. g. umbrella sampling method [4], which achieves the sampling of the overall space by adding so called “biasing potentials”. Adding biasing potentials have the same effect as changing temperatures, only written in a different term, and it can be seen as changing temperatures indirectly. The temperature ranges of sampling are not randomly chosen; based on different systems and different energy barriers, proper temperature ranges are required for efficient sampling; overlaps of the energy space among those different temperatures are important, too.

Another category of simulation methods focusing on capturing the transition state of a rare event with high barriers is called transition path sampling [5]. This method generates an ensemble of trajectories connecting the reactant to the product using Monte Carlo procedures called shooting and shifting. In an earlier paper [6] of ours, we tried to combine the umbrella sampling and transition path sampling together, and achieve more efficient sampling of the reaction paths. To represent earlier work done in the last two years, this paper is appended at the end of this thesis.

B. Research background

Many small animals, e.g., rotifers, tardigrades, nematodes, can survive dehydration for a long time; this suspended animation state is known as anhydrobiosis [7]. In anhydrobiosis, every metabolic process is switched off, which can be restored without any irreversible damage upon rehydration. A common feature for the animals during anhydrobiosis is that, there usually exist large amounts of sugar, particularly trehalose (figure 1), which is considered to be the most effective protection substance [8-11].

Based on this observation, many studies have been conducted on the protection effect of trehalose on different biomolecules, mainly proteins and membranes [12-17]. Three main hypotheses have been developed to explain the protection effect. The water-replacement hypothesis proposes that sugar can take the place of water molecules and form direct hydrogen bonds with the polar or charged groups of biomolecules, thereby stabilizing the native structure of proteins or phospholipid membranes in the absence of water [6, 18-21]. The water entrapment hypothesis, on the contrary, suggests that sugar molecules would concentrate residual water molecules close to the biomolecule, thus preserving the biostructure in a solvated position [22-24]. The third hypothesis, namely

the vitrification hypothesis, proposes that sugar molecules protect the biomolecular structure by forming amorphous glasses, which would reduce structural fluctuations and prevent denaturation or mechanical disruption [25-29]. However, these mechanisms are not necessarily exclusive to each other [9, 21, 26, 30-32]; vitrification can occur simultaneously with direct interactions of the biomolecule with sugars, or with entrapped residual water molecules around the biomolecule, depending on the type of the biomolecule.

In 2007, it was first reported by Bo Zhu et al. that trehalose can stabilize the three-dimensional structure of calf-thymus NaDNA (CT-DNA) in the dry state [33]. In this study, CT-DNA and trehalose were co-dissolved in deionized water, and desiccated at room temperature for two weeks until achieving a constant mass. The dehydrated samples were examined using the Fourier transform infrared (FTIR) technique. By analyzing the FTIR spectra, several conclusions were drawn: the CT-DNA persists in the B-form double-stranding in the dehydrated state; the protection effect is saturated at a 1:1 trehalose over nucleotide ratio; trehalose exists in an amorphous instead of a crystalline state in the dry sample. It was also proposed that “the stabilization effect of trehalose on DNA should basically arise from its ability to tightly hydrogen bond to phosphate groups of DNA, which leads to screening of the large phosphate-phosphate repulsion”.

To prove the protection of trehalose on DNA in the dry state, and also to find out how the trehalose molecules protect DNA, computer simulations would be very helpful. In the current study, we try to use Molecular Dynamics (MD) simulation to study the DNA-trehalose system in a relatively dry state, to provide an in-depth analysis of the atomic-level interactions that arise between trehalose molecules and a model decamer DNA in anhydrous environments.

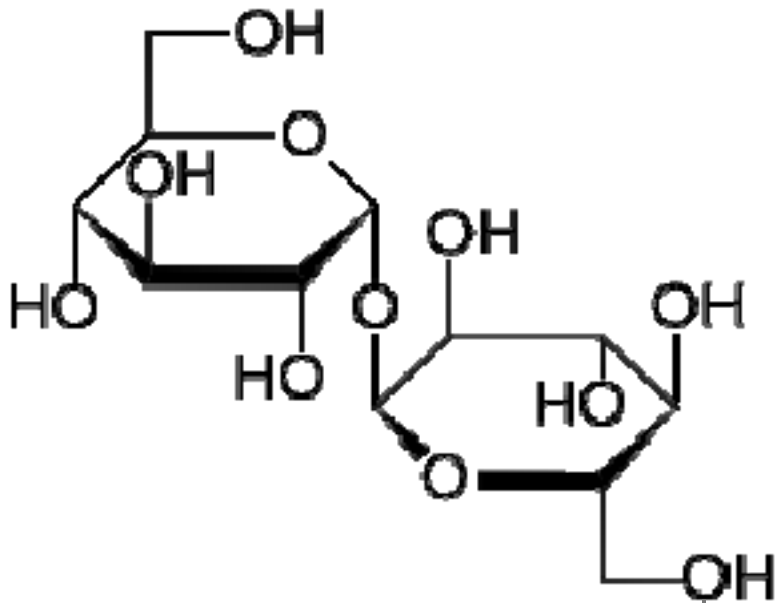

Figure 2. The molecular structure of trehalose.

II. METHODS

A. System description

A 53.7 Å × 48.8 Å × 75.7 Å crystalline trehalose dihydrate box was constructed, and a B-form DNA polyA-polyT decamer was solvated by the box (Figure 2). The system involves the decamer DNA, 18 sodium ions to neutralize the DNA, 407 trehalose molecules and 814 water molecules. A comparative system consists in a DNA polyG-polyC decamer solvated in a same trehalose dihydrate box, with the same number of sodium ion, trehalose and water molecules. The trehalose dihydrate box was built from the crystal structure unit provided by Cambridge Crystallographic Data Centre. Because of the regulated structure of trehalose dihydrate, there were cavities between the DNA and its surrounding trehalose dihydrate molecules when the system was first built (Figure 3a). The cavities can be filled (Figure 3b) by running 1ns simulation at 600 K with the DNA structure restricted (so as not to cause any irreversible damage to the double-strand DNA). Also the trehalose dihydrate structure became amorphous, to be the same with the experimental result.

B. Simulation details

Molecular dynamics simulations were performed using the AMBER 9 package. A three dimensional periodic boundary condition was applied. The particle mesh Ewald method was used to calculate the electrostatic energies and forces. The Van der Waals interactions and the real space part of the Ewald sum were truncated at 10Å. A time step of 1 fs was employed, and chemical bonds including H were constrained using the SHAKE algorithm. Two different temperatures, 300 K and 450 K were used in the simulation.

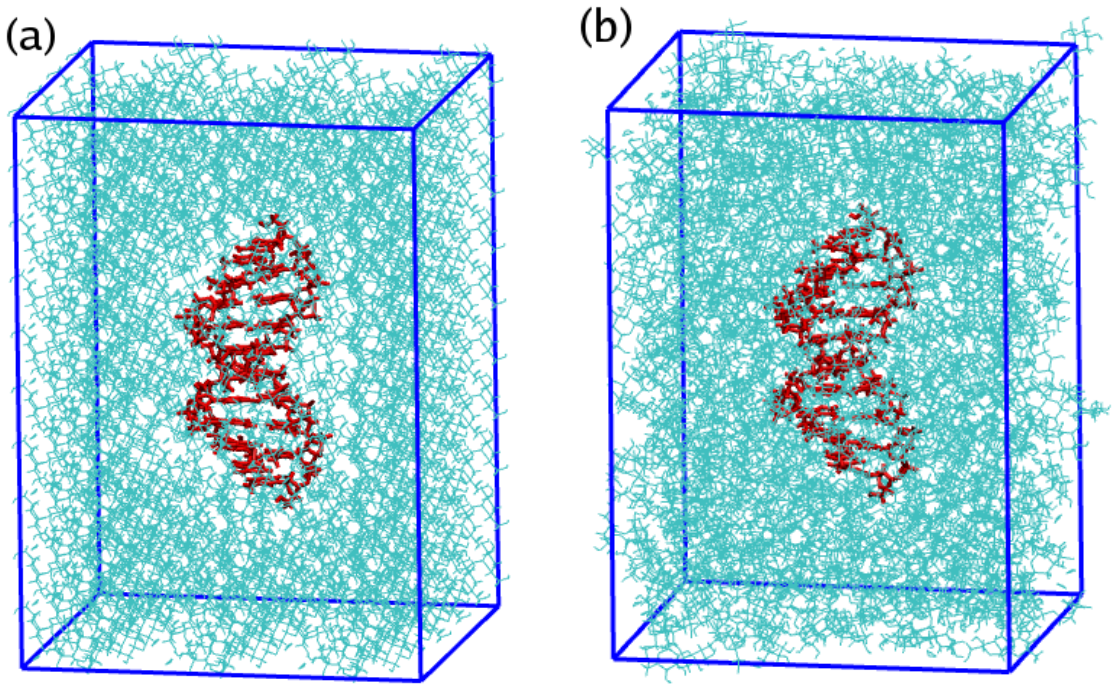

Figure 3. The simulation system. (a) initial structure (b) after 1 ns simulation at 600 K (The molecule shown in red depicts the 10-mer DNA, and the molecules shown in blue represent the trehalose and water molecules.)

For the 300 K simulation, because trehalose dihydrate (T_m= 370 K) exists in a solid state at this temperature, it would take very long time for the system to equilibrate. To overcome this difficulty, the system was first heated up to 600 K with the DNA structure fixed in its initial structure and run for 1ns, then cooled down to 300 K and run for another 1 ns (no configurational restriction on DNA at 300 K). Fifteen such heating up-cooling down cycles were conducted. Within each of the cycles, the 600 K part would allow trehalose and water molecules to move freely with a certain fixed DNA structure (the final structure from last 300 K cycle), then the 300 K part would allow the DNA to relax its structure to some extent under a rigid "solvent" condition obtained from the 600 K simulation. Considerably the system would gradually equilibrate, or at least optimize to a near equilibrium state. For the 450 K simulation, the system first experienced a 300 K- 600 K- 300 K cycle, so that the cavities between the DNA and trehalose molecules in the starting structure were filled, and then the system was run at 450 K for 20 ns. Constant volume simulations were performed during the heating up and cooling down processes, as well as in all 600 K simulations except during the first cycle (velocities will exceed the limit during heating up under constant pressure once the cavities have been filled). Constant pressure simulations were conducted for the 300 K and 450 K simulations.

C. Methods of analysis

In the structural analysis, the root mean square deviation (RMSD) was used as a standard to evaluate how the structures are compared to the initial structure, and it is defined as the minimum of the square root of the mean squared error (MSE). For a certain system containing n random vectors (atoms), the RMSD formula is

$$RMSD(\theta_1,\theta_2)=\sqrt{MSE(\theta_1,\theta_2)}=\sqrt{E((\theta_1-\theta_2)^2)}=\sqrt{\frac{\sum_1^n(x_{1,i}-x_{2,i})^2}{n}} \tag{7}$$

where θ_1 and θ_2 indicates the two structures compared.

Hydrogen bonds were monitored and classified according to the pairs of species present in the simulations (trehalose-DNA, and water-DNA). The criterion used to define a hydrogen bond was a hydrogen-acceptor distance shorter than 0.35 nm and a donor-hydrogen-acceptor angle larger than 120˚ [34]. For the DNA, all the oxygen atoms in the phosphate group of DNA, two nitrogen atoms on adenine, and two oxygen atoms on thymine were considered as electron donors, and one hydrogen atom on adenine group was treated as an electron accepter. For the trehalose, all the eleven oxygen atoms were electron donors and all the eight hydrogen atoms within the OH groups were accepters. The hydrogen bonds formed by one pair of species were further classifed into two categories: hydrogen bonds formed by the backbone DNA (phosphate groups) and those formed by the basepairs of DNA.

Diffusion constants (D) were calculated for trehalose and water molecules. These were obtained form the Einstein relation:

$$2NDt = < \left| R(t) - R(0) \right|^2 > \tag{8}$$

where the right-hand side of the equation refers to the mean-square positional displacement of a specific atom, and N is the number of translational degrees of freedom considered. For most simulation systems like our simulation system, N is simply equal to three.

III. RESULTS AND DISCUSSION

A. Structure analysis

The structures of the polyA-polyT DNA during the simulation were compared with the initial B-form structure using molecular visualization software VMD. Within the 600 K- 300 K simulation cycles, only the 300 K trajectories (15ns altogether) were used for analysis, since the DNA structures were fixed at 600 K. The root mean square deviation (RMSD) values for all heavy atoms of DNA in the 300 K simulation are shown in Figure 4. The RMSD value had a sharp increase at the beginning, and then gradually stabilized below 2 Å. The 300 K simulation converged very slowly, since the structures of DNA and trehalose were not allowed to optimize together: the DNA structure had to be constrained at 600 K in order not to be irreversibly damaged, and at 300 K the trehalose molecules were highly immobile. Until the end of the 15th cycle, the RMSD value still seemed to be increasing. Therefore, instead of continuing the cycles, we conducted 20ns simulation at 450 K. Trehalose dihydrate would exist in a liquid state at this temperature, so the DNA configuration was able to optimize together with the "solvent" molecules. It was very surprising that the RMSD values of DNA stabilized around 3 Å after 3ns during the 20ns simulation (Fig. 5), indicating that the B-form DNA structure can still be well protected even at such a high temperature. A comparative study of this 10-mer polyA-polyT DNA in water at 450 K showed that the DNA structure was totally disrupted after only 5 ns simulation. To study whether the protection effect was specific for A-T base pairs, a same 300 K simulation was conducted on a system consisting of a 10-mer polyG-polyC DNA. The pattern of the RMSD values for this DNA showed no difference from that of the polyA-polyT DNA (Fig. 4).

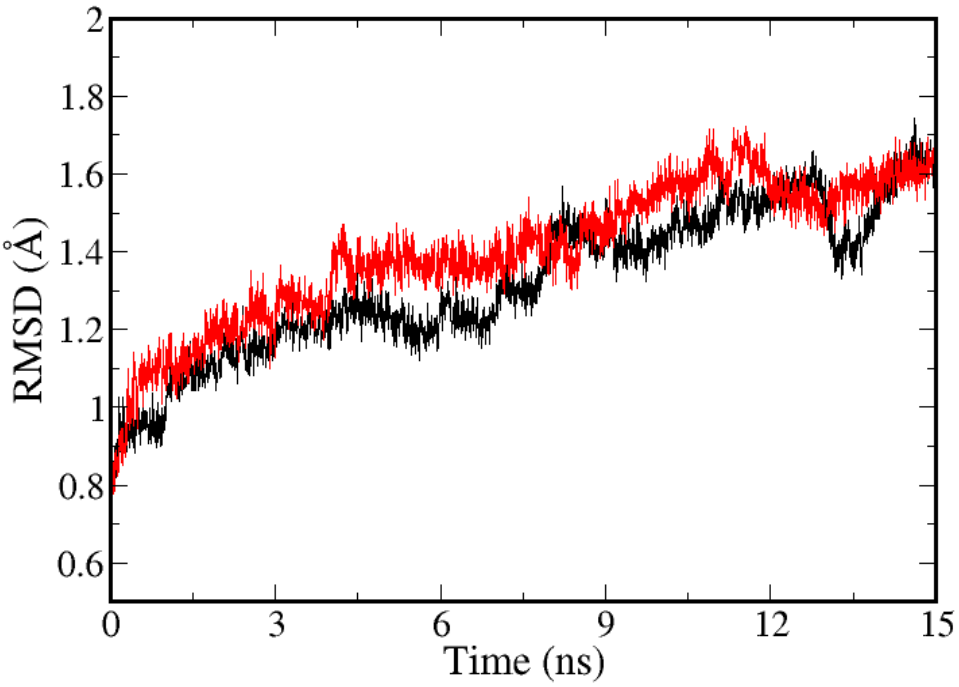

Figure 4. RMSD values for all heavy atoms of DNA at 300K. (Black: polyA-polyT DNA; red: polyG-polyC DNA.)

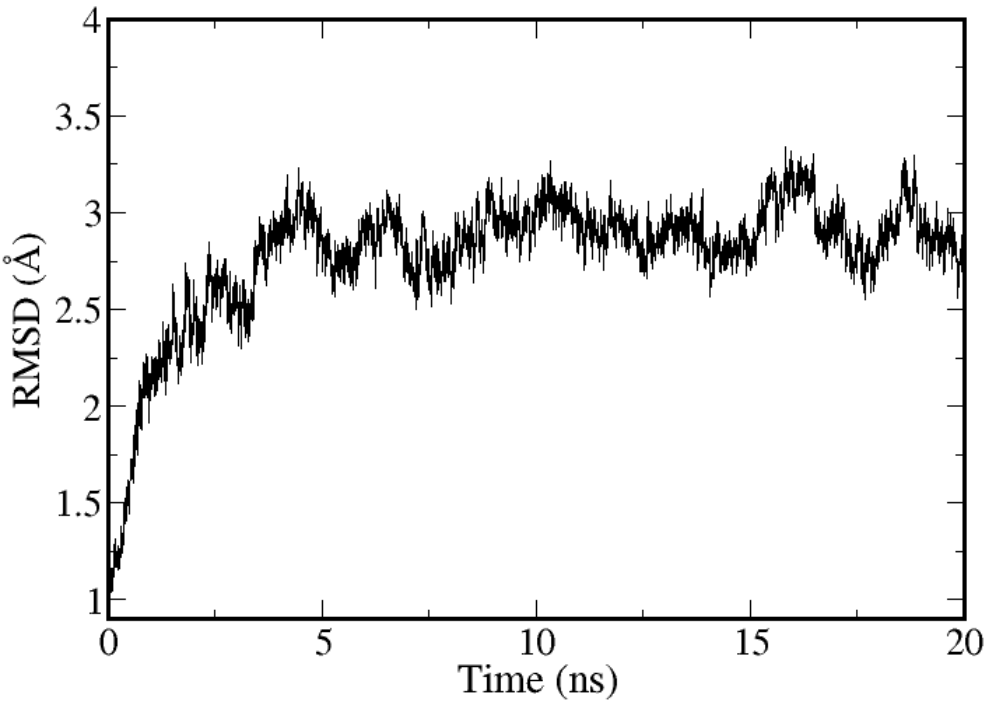

Figure 5. RMSD values for all heavy atoms of DNA at 450 K. (polyA-polyT)

B. Hydrogen bond analysis

The hydrogen bonds formed between trehalose and DNA and between water and DNA were analyzed. Only the eight base pairs with their phosphate groups in the middle of the DNA were used for this analysis. Hydrogen bonds can be formed in a number of ways between DNA and trehalose or water. For the adenine-thymine base-paired DNA molecule, the four oxygen atoms within each of the 16 phosphate groups, the two unsaturated nitrogen atoms on each adenine group, and the two oxygen atoms on each thymine group, can serve as electron donors; one hydrogen atom (connected to N) within the adenine group can serve as an accepter. For a trehalose molecule, all the eleven oxygen atoms can serve as donors, and all the eight hydrogen atoms within the OH groups can serve as accepters. The hydrogen bonds were analyzed between donors/accepters from DNA and accepters/donors from trehalose or water. For the convenience of discussion, we classified the hydrogen bonds into two categories: those formed with backbone DNA (the phosphate groups), and those formed with base pairs. Hydrogen bonds in this work were defined using the criteria suggested by Brady and Schmidt [28] (the oxygen-oxygen distance should be < 3.5 Å and the $O-H\cdots O$ angle > 120˚.

For the 300 K simulation, the number of hydrogen bonds between backbone DNA and trehalose stabilized at around 65, and the number between base pairs and trehalose was around 20 (Fig. 6). The number of hydrogen bonds formed between base pairs and trehalose increased slightly during the 15 cycles. This increase is likely due to the structural relaxation of the DNA, which slightly tilted the base pairs allowing them to be more available to form hydrogen bonds. The number of hydrogen bonds formed by water with the backbone DNA and with base pairs stayed around 35 and 20, respectively (Fig.

6). From these numbers, we can see that hydrogen bonds formed between trehalose and backbone DNA constitute the majority of hydrogen bonds, in accordance with the prediction by Bo Zhu et. al. that the main protection effect is due to trehalose hydrogen bonding to the phosphate groups of DNA. Trehalose can also interact with other polar groups of DNA, although with a much smaller number of hydrogen bonds formed. The water molecules were not excluded from the surface of DNA upon trehalose-DNA interactions. To compare the ability of forming hydrogen bonds with DNA, a simple argument is applied: each trehalose molecule has 8 hydrogen atoms which can form hydrogen bonds with the phosphate groups; each water molecule has 2; the molecule number ratio of trehalose: water=1:2, so assuming that water and trehalose are equally favorable for hydrogen bonding with DNA, trehalose will form twice as many hydrogen bonds as water. Since the hydrogen bond ratio from our simulation is 65:35, which is a little less than 2:1, trehalose does not seem more favorable than water to form hydrogen bonds with DNA. However, this is not conclusive since configurational restrictions would limit the maximum bonds to be formed between the trehalose and DNA.

At 450 K, the hydrogen bond numbers showed a similar pattern to that of the 300 K simulation result. The number of hydrogen bonds between trehalose and backbone/base pair DNA stabilized at around 60 and 13, respectively; that between water and backbone/base pair DNA stabilized at around 30 and 18, respectively (Fig. 7). The largest contribution of hydrogen bonds was still from trehalose and backbone DNA. But the hydrogen bond numbers of all the four types were lower than the corresponding 300 K results. This is because at a higher temperature, the system has a higher energy, resulting in fewer numbers of hydrogen bonds. Also, the ratio of hydrogen bonds formed by trehalose with the backbone DNA to water with the backbone DNA increased to 60:

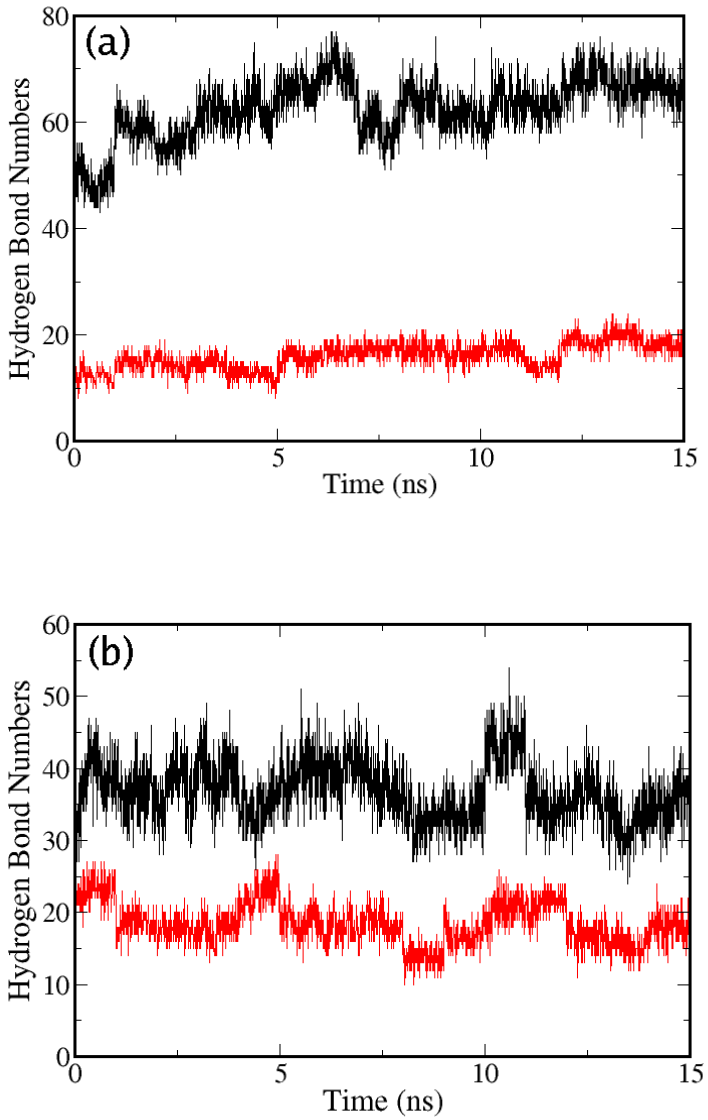

Figure 6. Number of hydrogen bonds at 300K. (a) formed by trehalose and backbones/base pairs of DNA (black/red); (b) formed by water and backbones/base pairs of DNA (black/red).

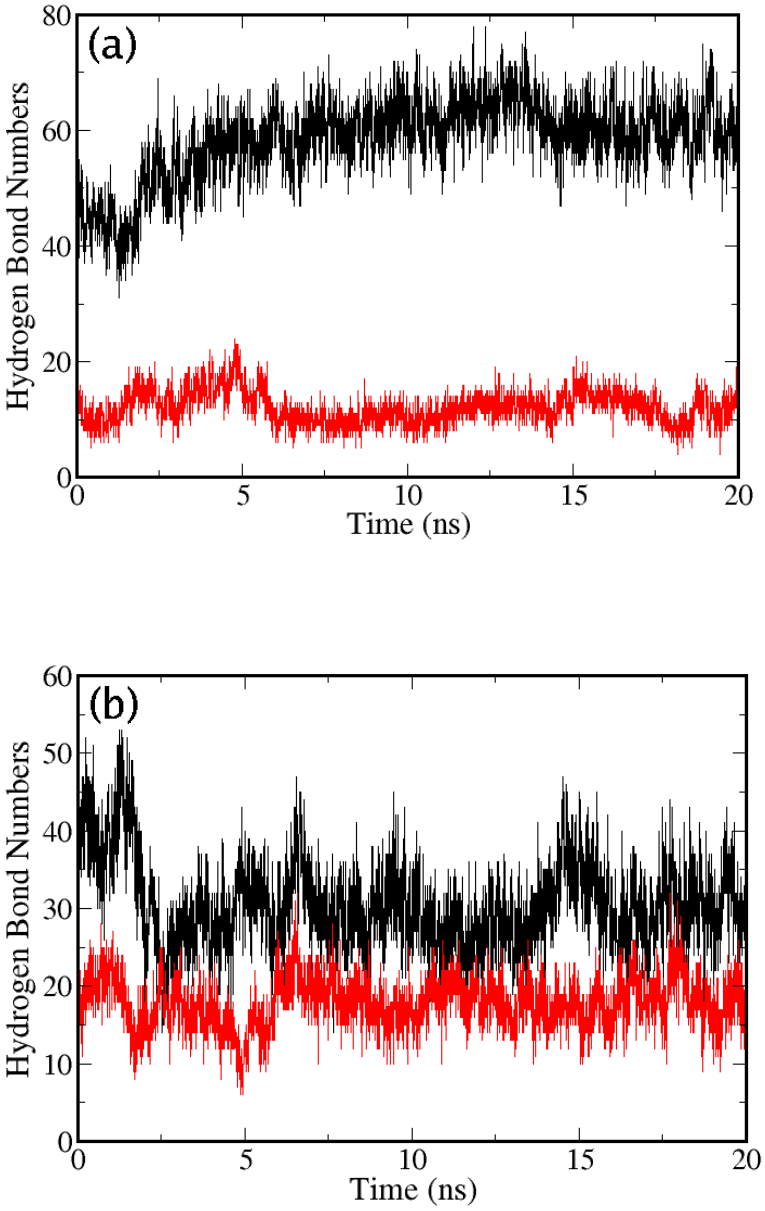

Figure 7. Number of hydrogen bonds at 450K. (a) formed by trehalose and backbones/base pairs of DNA (black/red); (b) formed by water and backbones/base pairs of DNA (black/red).

30, indicating a greater importance of trehalose in protecting DNA at higher temperatures. The 450 K simulation has a good convergence in both DNA structure and hydrogen bond numbers. Although the 300 K simulation seemed not to converge, the results are verified since they agree with the 450 K simulation.

To find out which of the eight OH groups within a trehalose molecule has the most propensity to form hydrogen bonds with the backbone DNA, we analyzed the hydrogen bond numbers formed by each OH group. Each trehalose molecule consists of two glucoses (namely 0GA and 1GA in Amber) linked by the 1-1 α bond, and the four OH groups within each glucose are named OH2, OH3, OH4, and OH6, by the order of the carbon atoms connected to them (Fig. 1). By analyzing the 20ns 450K simulation trajectory, the number of hydrogen bonds formed by the OH groups were found to be in the order of OH3 > OH4 > OH2 > OH6 for both 0GA and 1GA (Fig. 8). This order is in agreement with an earlier theoretical study of trehalose- phospholipids interactions [11].

Next we were asking whether the OH groups have correlations in forming hydrogen bonds with the backbone DNA. Only weak positive correlations among neighboring OH groups were found for the trehalose. This should purely be due to the configuration of the trehalose: if one OH groups is close to the phosphate groups of DNA and can form hydrogen bonds, then a neighboring OH is more likely to be close to the DNA too.

Since the analysis of hydrogen bond numbers provides limited information on how exactly trehalose and DNA interact with each other, we output forty glucose residues nearest to the DNA and took a close look (Fig. 9). Snapshots at both 300K and 450K were examined, and several typical kinds of interactions were recognized (Fig. 10-11). Over 90% of trehalose molecules would form hydrogen bonds with DNA through only

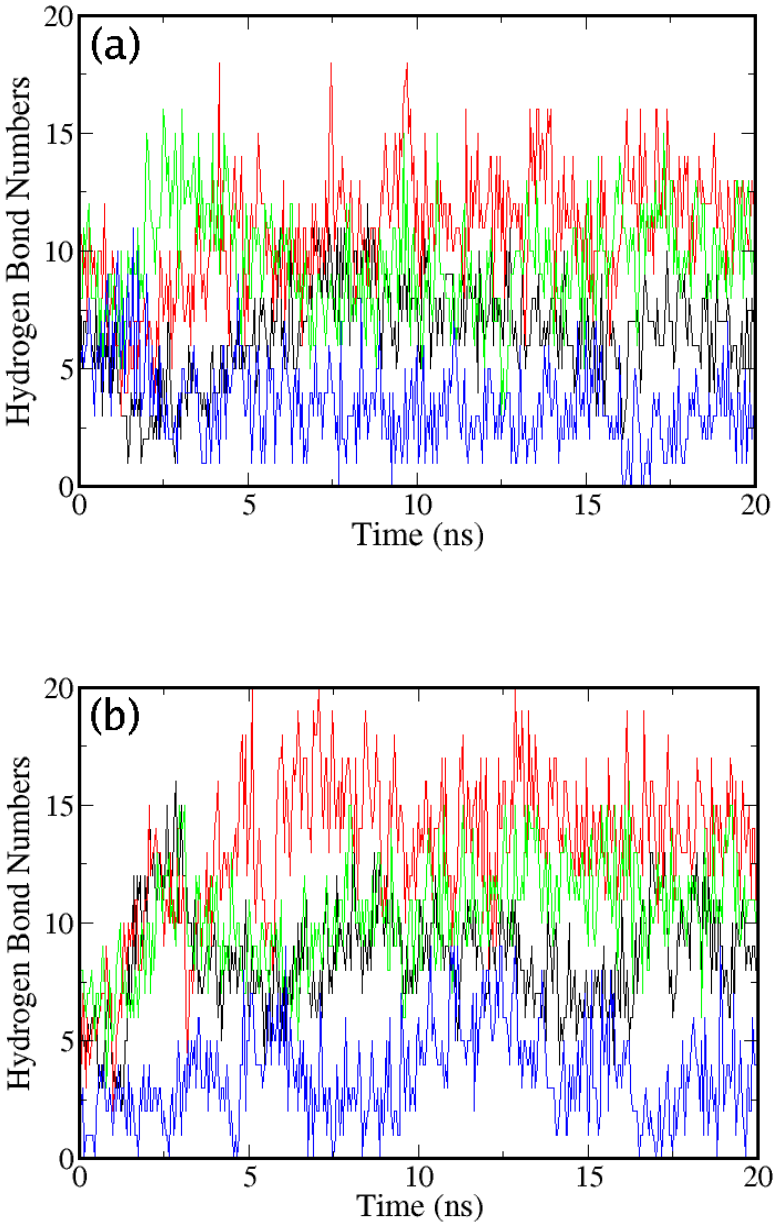

Figure 8. Number of hydrogen bonds formed by each OH group. (a) 0GA and (b) 1GA at 450K. (Black: OH2, Red: OH3, Green: OH4, Blue: OH6)

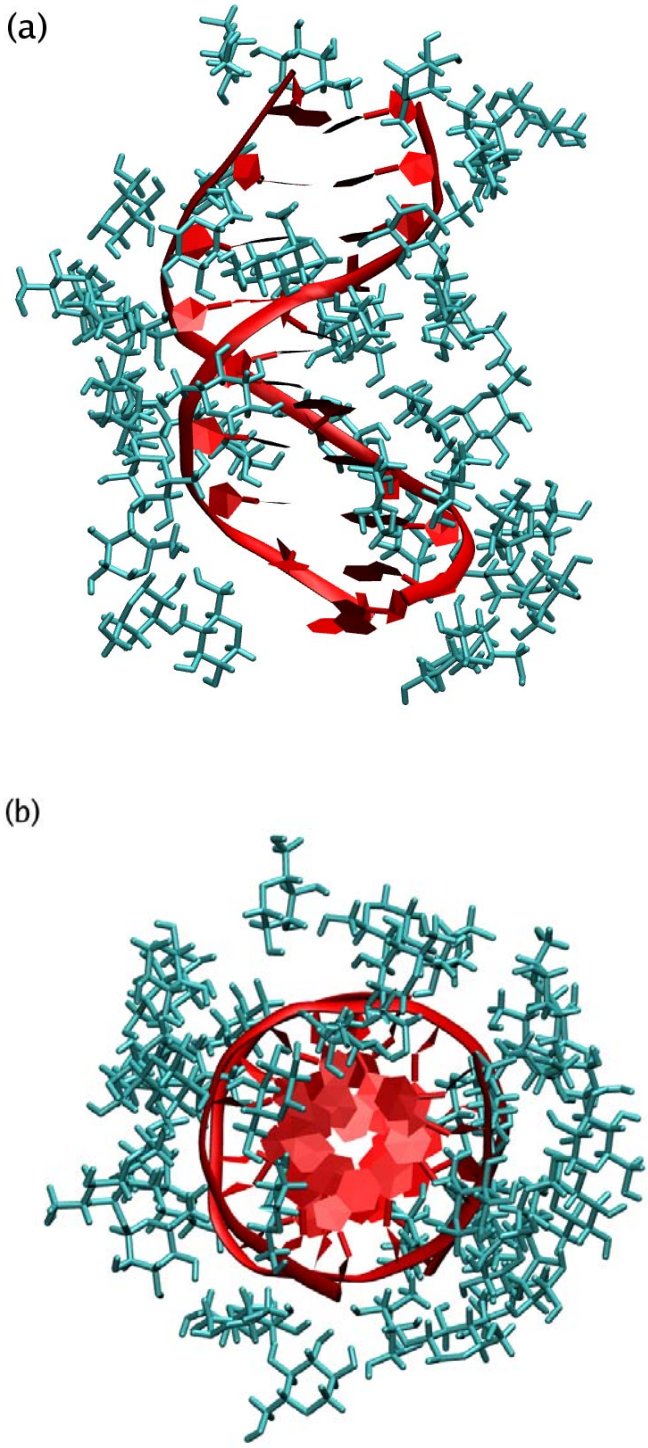

Figure 9. Forty closest glucose residues of trehalose and the decamer DNA. (a) side view; (b) top view.

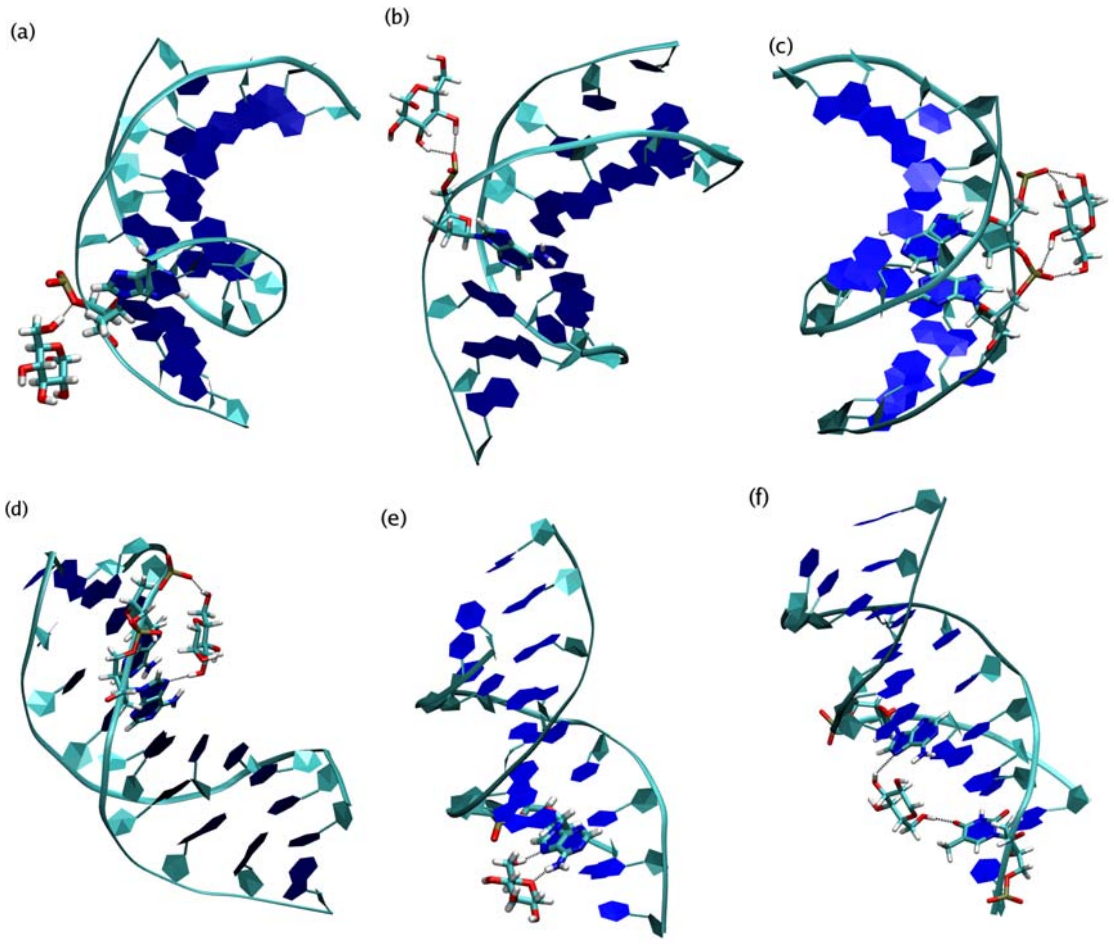

Figure 10. Trehalose interacts with DNA through one glucose. (the other glucose does not form hydrogen bonds with DNA and thus not shown) (a) by forming one bond; (b) by forming multi bonds with one phosphate groups; (c) with two phosphate groups; (d) with one phosphate group and one base pair; (e) with two base pairs; (f) with one base pair.

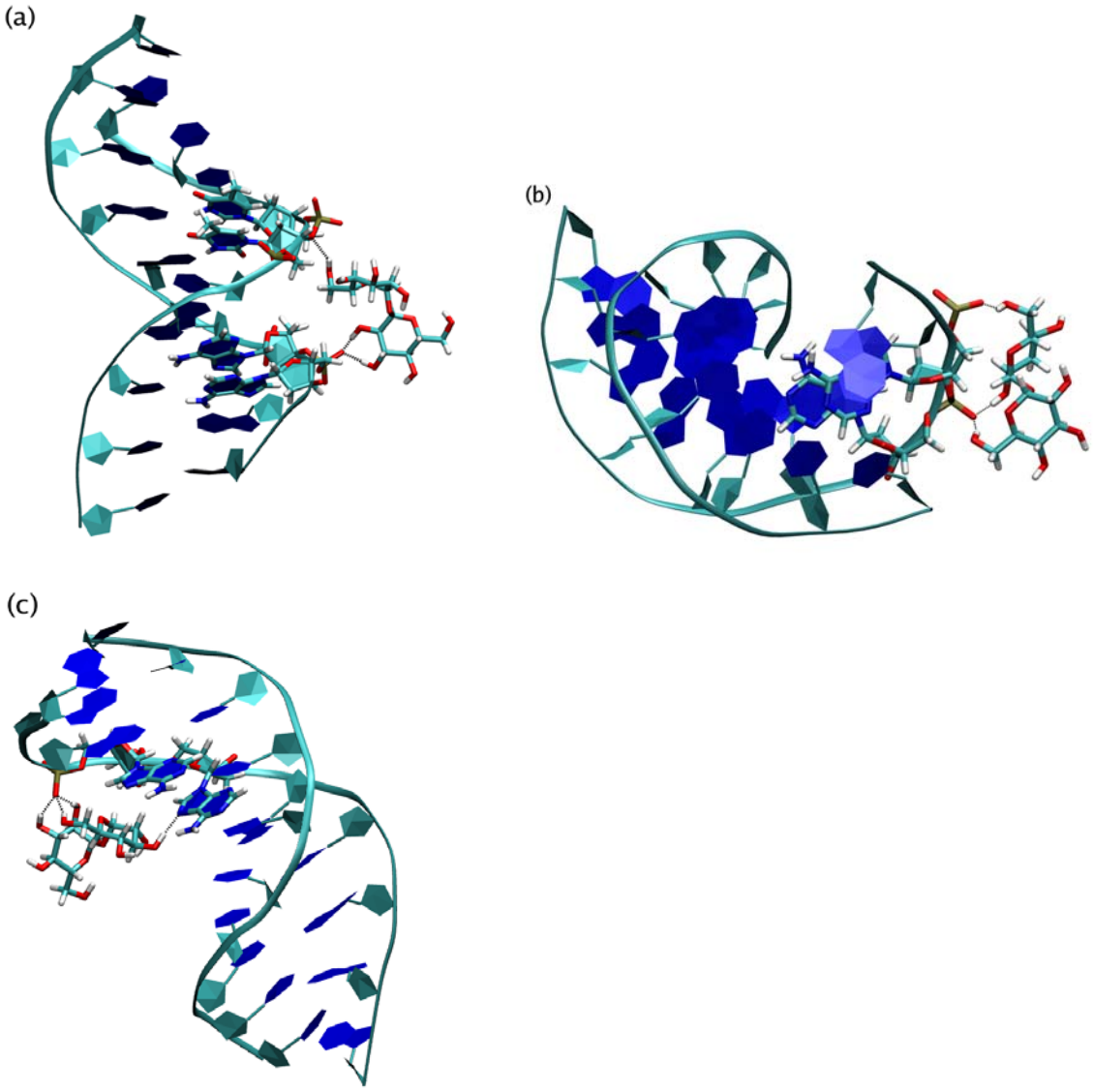

Figure 11. Trehalose interacts with DNA through both glucoses. (a) by forming hydrogen bonds with two phosphate groups on different strands; (b) with two phosphate groups on one strand; (c) with one phosphate group and one base pair.

one glucose (Fig. 9). Within this class, around 1/3 of the glucoses form only one hydrogen bond with the DNA (Fig. 9a), while the rest glucoses would form more than one hydrogen bonds: they can form multi bonds with DNA phosphate groups (Fig. 9b, c), or with base pairs (Fig. 9d, e), or with both a phosphate group and a base pair (Fig. 9f). Only less than 10% of trehaloses can form hydrogen bonds through both glucoses (Fig. 10). Within this class, a trehalose molecule can usually form more than two hydrogen bonds with the DNA: they can bind to two different phosphate groups, either in one strand (Fig. 11a), or in two strands (Fig. 10b); they can also form hydrogen bonds with both a phosphate group and a base pair (Fig. 10c). A frequently observed feature was that two hydrogen atoms from two neighboring OH groups bond to a same phosphate oxygen (Fig. 9b, c, Fig. 10a, b, c). This observation provides a more direct proof of the weak coupling between neighboring OH groups. But in general, there was no dominant pattern of how trehalose would form hydrogen bonds with DNA. It should be the strong network of hydrogen bonds formed by trehalose that protects the DNA structure. The trehalose molecules not only interact with the DNA, but also form a large number of hydrogen bonds among themselves, constituting a glassy/viscose medium, thereby reducing structural fluctuations of DNA and preventing its denaturation. This statement very well corresponds to the vitrification hypothesis originally proposed for trehalose protecting proteins and membranes.

A diffusion constant analysis on water and trehalose provided further proof for the vitrification mechanism. At 300 K, the diffusion constants for water and trehalose were calculated to be 1.67×10^{-7} cm^2 s^{-1} and 1.17×10^{-8} cm^2 s^{-1}, respectively; while at 450 K, these two values were found to be 4.69×10^{-6} cm^2 s^{-1} and 9.62×10^{-8} cm^2 s^{-1}, respectively. All these values are much smaller than the diffusion constant of water at room

temperature (1.67×10^{-5} cm^2 s^{-1}).

IV. CONCLUSIONS

Our MD simulation results supported the experimental observation that the B-form structure of DNA was well protected by trehalose in the dry state, and also suggested that this protection can be effective at high temperatures. Trehaose molecules formed direct hydrogen bonds with DNA while water molecules were neither excluded from interacting with DNA nor concentrate water close to the DNA. Since trehalose molecules could be seen as replacing waters around the DNA, the water-replacement hypothesis was suitable for the DNA-trehalose system. Hydrogen bonds formed between trehalose and phosphate groups of DNA was the major factor of the protection effect. Trehalose interacted with DNA in a random pattern by forming a glassy intermediate. Through structure and diffusion constant analysis, it was believed that the vitrification hypothesis provided a good description of how trehalose protects DNA.

REFERENCES

1. J. W. Ponder, D. A. Case, *Prot. Chem.* **66**, 27 (2003).
2. M. P. Allen, D. J. Tildesley, Computer Simulation of Liquids. Oxford University Press, Oxford, UK. 78-82 (1987).
3. H. Nymeyer, S. Gnanakaran, A. E. Garcia, *Meth. Enzymol.*, **383**, 119 (2004).
4. G. M. Torrie, J. P. Valleau, *J. Comp. Phys.* **23**, 187 (1977).
5. C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, *J. Chem. Phys.* **108**, 1964 (1998).
6. X. Fu, L. Yang, Y. Q. Gao, *J. Chem. Phys.* **127**, 154106 (2007).
7. J. S. Clegg, *Comp. Biochem. Physio.* **14**, 135 (1965).
8. L. M. Crowe, D. S. Reid, J. H. Crowe, *Biophys. J.* **71**, 2087 (1996).
9. J. H. Crowe, J. F. Carpenter, L. M. Crowe, *Annu. Rev. Physiol.* **60**, 73 (1998).
10. J. F. Carpenter, J. H. Crowe, *Biochemistry* **28**, 3916 (1989).
11. A. M. Massari, I. J. Finkelstein, B. L. McClain, A. Goj, X. Wen, K. L. Bren, R. F. Loring, M. D. Fayer, *J. Am. Chem. Soc.* **127**, 14279 (2005).
12. J. H. Crowe, L. M. Crowe, D. Chapman, *Science* **223**, 701 (1984).
13. J. H. Crowe, L. M. Crowe, J. F. Carpenter. C. A. Wistrom, *Biochem. J.* **242**, 1 (1987).
14. J. H. Crowe, F. Tablin, W. F. Wolkers, K. Gousset, N. M. Tsvetkova, J. Ricker, *Chem. Phys. Lipids.* **122**, 41 (2003).
15. P. Ballone, M. Marchi, C. Branca, S. Magazu, *J. Phys. Chem. B.* **104**, 6313 (2000).
16. F. Sussich, C. Skopec, J. Brady, A. Cesaro, *Carbohydr. Res.* **334**, 165 (2001).
17. C. S. Pereira, R. D. Lins, I. Chandrasekhar, L. C. G. Freitas, P. H. Hunenberger, *Biophys. J.* **86**, 2273 (2004).

18. J. H. Crowe, J. S. Clegg, and L. M. Crowe, the Properties of Water in Foods (ISOPOW 6). Chapman and Hall, New York. 440-455 (1998).

19. J. H. Crowe, L. M. Crowe, J. F. Carpenter, S. Prestrelski, F. A. Hoekstra, P. de Araujo, and A. D. Panek. Handbook of Physiology, Vol. 2. Oxford University Press, Oxford, UK. 1445-1477 (1997).

20. J. H. Crowe, F. A. Hoekstra, and L. M. Crowe, *Annu. Rev. Physiol.* **54**, 579 (1992).

21. J. F. Carpenter, S. J. Prestrelski, T. J. Anchordoguy, and T. Arakawa, Formulation and Delivery of Proteins and Peptides, American Chemical Society, Washinton, DC. 134-147 (1994).

22. P. S. Belton, A. H. Gil, *Biopolymers.* **34**, 957 (1994).

23. G. Cottone, G. Ciccotti, and L. Cordone, *J. Chem. Phys.* **117**, 9862 (2002).

24. R. D. Lins, C. S. Pereira, and P. H. Hunenberger, *Proteins.* **55**, 177 (2004).

25. W. Q. Sun, T. C. Irving, and A. C. Leopold, *Physiol. Plantarum.* **90**, 621 (1994).

26. W. Q. Sun, A. C. Leopold, *Comp. Biochem. Physiol.* **117**, 327 (1997).

27. W. Q. Sun, A. C. Leopold, *Ann. Bot*. **74**, 601 (1994).

28. W. Q. Sun, A. C. Leopold, L. M. Crowe, and J. H. Crowe, *Biophys. J.* **70**, 1769 (1996).

29. R. J. Williams, A. C. Leopold, *Plant Physiol.* **89**, 977 (1989).

30. J. H. Crowe, A. E. Oliver, N. Tsvetkova, W. Wolkers, and F. Tablin, *Cryobiology.* **43**, 89 (2002).

31. L. M. Crowe, *Comp. Biochem. Physiol.* **131**, 505 (2002).

32. J. S. Clegg, *Comp. Biochem. Physiol.* **128**, 613 (2002).

33. B. Zhu, T. Furuki, R. Okuda, M. Sakurai, *J. Phys. Chem. B Lett.* **111**, 5542 (2007).

34. J. W. Brady, R. K. Schmidt, *J. Chem. Phys.* **81**, 3684 (1993).

APPENDIX

Selective Sampling of Transition Paths*

Xuebing Fu[1], Lijiang Yang[1] and Yi Qin Gao♦

Department of Chemistry
Texas A&M University, College Station, TX 77843

Abstract

In this short paper, we introduce an approximate method for the quick estimate of rate constants based on a simple sampling method of reactive transition paths over high energy barriers. It makes use of the previously introduced accelerated molecular dynamics (MD) simulation method to generate initial points for trajectory shooting. The accelerated MD simulations, although with the loss of real dynamics, lead to a quick calculation of thermodynamic properties and at the same time produce an ensemble of configurations with an enhanced sampling over the phase space that is more "reactive". The forward/backward trajectory shooting as that used in the transition path sampling method is then initiated from the configurations obtained from accelerated MD simulations to generate transition paths on the original unbiased potential. This method selectively enhances sampling of successful trajectories and at the same time accelerates significantly the calculation of rate constants.

1. Introduction

During the last several decades, a lot of efforts have been dedicated to improving

---

* Reprinted with permission from Fu, X., Yang L., Gao, Y. Q. Selective sampling of transition paths, *Journal of Chemical Physics* 127, 154106, 2007.

[1] These authors contributed equally to this work

♦To whom correspondence should be addressed. E-mail: yiqin@mail.chem.tamu.edu

computational efficiency so that thermodynamics and/or dynamics of complex systems can be calculated at low costs. Many of the systems under study involve a large number of particles and a wide distribution of time and length scales, and computations are being challenged to approach longer time scales. A variety of methods have been developed to enhance sampling over the phase space and thus allow fast calculations of thermodynamics properties. Methods such as J-walking,[1] adaptive umbrella sampling,[2] simulated tempering,[3,4] parallel tempering,[5-8] catalytic tempering,[9] replica exchange,[10,11] multi-canonical simulations,[12,13] metadynamics,[14] conformational flooding,[15] conformational space annealing,[16] hyperdynamics,[17,18] potential smoothing methods,[19] stochastic potential switching[20] and the approximate potential method[21] belong to this category. A straightforward way of adding a bias potential to the potential energy function of a complex system to enhance sampling of the important but rarely sampled high energy states has already been suggested in the original paper by Torrier and Valleau.[22] Recently, in a series of publications Hamelberg *et al.*,[23-26] proposed and applied a method to alter the potential energy landscape, leading to an efficient sampling of the conformational space. In this method the sampling over the high energy range was increased at the cost of a largely reduced sampling over the states with lower energies. More recently, we proposed an approach to accelerate MD simulations in which a bias potential was added to control the enhanced sampling in a desired range of energy (e.g., near the energy barrier and in the entire energy range that is important at the given temperature) without under sampling the low energy or over sampling the high energy range, leading to a more uniform distribution as a function of energy. The method was shown to be efficient in free energy simulations[27] as well as in folding a protein.[28] One of the common problems of the methods mentioned above is that the real dynamics of the

system is lost in the accelerated simulations.

The well-celebrated transition path sampling (TPS) method, developed by Chandler and coworkers,[29-32] on the other hand, focuses on enhancing the sampling of transition paths, and has been shown to be effective in calculating the dynamics of slow processes.[33,34] This method generates an ensemble of trajectories connecting the reactant to the product using Monte Carlo (MC) procedures called shooting and shifting. Two difficulties arise in the application of TPS. Firstly, the calculation of rate constant is very computer time consuming and secondly it is often difficult to generate the initial trajectory from which the shifting and shooting procedure starts, when the system is complex. To overcome the first difficulty, a more efficient transition interface sampling (TIS) method[35,36] which is based on the effective positive flux formalism was developed. A related partial path TIS (PPTIS) method,[37] a variation of TIS that samples shorter paths was also introduced. TIS permits relatively fast calculations of rate constants but does not generate continuous trajectories connecting the reactant to the product and thus more detailed dynamics information is lost in the application of this method.[38] On the other hand, in the application of TPS, several different approaches, for example, to initiate trajectories from the potential energy transition state,[39] or from a high temperature,[33] or targeted molecular dynamics simulations,[40] have been used to generate the initial trajectory, which is subsequently relaxed and used in the sampling for unbiased transition paths. These methods were all shown efficient in initiating the transition path sampling, although the subsequent rate constant calculation is still slow.

In this short paper, we present a selective sampling of transition paths method which

takes advantages of both accelerated MD simulation and transition path sampling (shooting) methods. As we show in the paper, since the former allows a fast sampling of the configuration space and the latter allows quick identification of transition paths, dynamics calculations (such as that for the rate constant and reaction pathways) can be accelerated when ideas of the two methods are combined. The rest of the paper is organized as follows: In Section 2, the theory and computational method are described in detail. The method is then applied to study the conformational changes of a simple system: methyl maltoside in Section 3. The results are then discussed in Section 4.

2. Computational Method

2.1 *Accelerated MD simulation*

When a high energy barrier is present, the chemical reactions are usually slow and difficult to be calculated using standard MD simulation method, in particular for a large system. In the accelerated MD simulation (umbrella sampling) a bias potential is added to the original potential to enhance the sampling of high energy ranges and thus allows more frequent transitions. As we showed earlier,[27,28] a desirable bias potential should help to generate a largely uniform distribution in and only in the entire important energy range, which includes the energies of states that are highly populated at the desired temperature as well as energies that are high enough for the reaction to occur. The over-sampling of other energy ranges (such as very high energy ranges) should be avoided. The bias potential can be written as a sum of Gaussian terms:[28]

$$f(V) = \sum_{i}^{n} a_i e^{-[(V-V_i)/\sigma_i]^2} \qquad (1)$$

where V_i is the energy of a state that is to be sampled with enhancement, a_i is a

negative constant which defines how much the enhancement will be. The trial and error strategy is used to determine a suitable bias potential. Firstly, a short trajectory (typically between 10 and 100 ps) is obtained using the standard MD simulation at the desired temperature. Then the trajectory is analyzed to yield information on the potential energy distribution. Two energy regions thus can be identified. One is the region that mostly sampled in the normal MD simulation. The sampling of this region should be decreased by adding a positive or close to zero bias potential. The second region is the low energy region which is not sampled enough but might be important. A negative bias potential should be added in this region to increase the sampling here. Similarly, a short standard MD simulation trajectory at a higher temperature (the high temperature is determined by how much one wants to accelerate the chemical transitions, typically 400K – 600K) is also obtained and analyzed. The mostly sampled energy region at this temperature is also important and a negative bias potential should be introduced here to increase the sampling. After identifying the three key energy regions, a bias potential like Eq. (1) is first added in the three key energy regions. And more Gaussian terms would be added to make the bias potential smooth (this is required by avoiding the introduction of artificial energy barriers). Another important issue is the magnitude of the bias potential. Generally, it can not be too large. This is because: (1) if the bias potential is too large, structures of molecules defined under current molecular mechanics force fields would be totally destroyed. (2) If the bias potential is too large, it is very difficult to obtain converged results in free energy and rate constant calculations. Certainly, the bias energy can not be too small either otherwise there would not be enough energy to drive the system to cross energy barrier. Since we have obtained energy distributions at room temperature and a higher temperature, densities of states at room temperature $\rho_1(V)$

and that at a higher temperature $\rho_2(V)$ are fitted roughly to a function $\rho(V)$ using a polynomial. If V_1 is an energy at which the bias potential equals zero and V_2 is an energy at which the magnitude of the bias potential is the largest, $\Delta V = (V_2 - V_1) - RT \ln[\rho(V_1)/\rho(V_2)]$ is used as the initial estimation of the depth of the bias potential. Finally, the bias potential is added to the molecular system to obtain a short trial trajectory at room temperature. A typical criterion we used is that in a short trajectory (e.g., ~ 10 ps) both low and high temperature ranges (see above) should be visited. If the energy range is broadened to desirable high and low energy ranges, the bias potential is considered to be a suitable one and can be used in following accelerated MD simulations. Otherwise parameters in Eq. (1) need to be adjusted or additional terms should be introduced according to the energy distributions of the two standard MD simulations and the current trial simulation with the bias potential. After several rounds of this trial and error procedure, a suitable bias potential should be determined. Although this strategy is not entirely quantitative, it is easy to implement and effective in practice. After the bias potential being determined, standard MD simulations will then be run with the modified forces (the sum of the original force *F* and the additional force $\frac{\partial f}{\partial V} \times F$). The trajectories are then used to calculate thermodynamic properties of the original (unbiased) system using standard recovering procedures.[27,28]

2.2. *Transition Path Shooting*

As discussed earlier, MD simulations performed using a biased potential allows fast convergence for the sampling of the configuration space and since the simulations are performed at a constant temperature, the momentums follow a Maxwell distribution.

Although using the bias potential method, one can easily generate successful trajectories connecting different states without the requirement of knowing the reaction coordinate, since these trajectories are obtained using a biased potential, they do not represent the transition paths for the original system. In the following a systematic method of selectively generating reactive transition paths for the original unbiased system is proposed. The basic assumptions of this method are: (1) the convergence in the free energy calculation indicates that a correct Boltzmann distribution in the configuration space has been achieved when they are weighted by the corresponding bias potential term.[27] (2) The total energy of the system is kept constant while shooting trajectories.[41] As a result of the above two assumptions, (3), trajectories generated on the *real* potential energy surface through constant energy shootings, using the phase space points sampled for the *biased* system as initial points, constitute an ensemble of *real* and *unbiased* transition paths for the *real* system when the initial points are weighted by their correct Boltzmann distribution. The trajectories generated from the biased potential certainly do not represent real trajectories of the real system. While in the current method, the points along these *biased* trajectories that are separated by a relatively large time interval (e.g. 100 fs) are considered as *independent* phase space points of the *real* system, although, their expected visiting probability in the real system is different by a factor of $\exp[f(V(r))/k_BT]$ from that observed in the biased system, as a result of the bias potential. Therefore, trajectories initiated from these phase space points on the original potential energy surface represent real trajectories for the real system (except for the coupling between the system and its surroundings not being precise, see the note[41] for the assumption (2)), and their probabilities of occurrence in the real system are recovered when the weighting factor for the initial points, $\exp[f(V(r))/k_BT]$, are taken into

account, under the assumption (1). And further, (4), the trajectories obtained using the bias potential provide a biased sampling of the phase space, from which the most "reactive" phase space region, which contains phase space points more likely to be on a reactive trajectory of a given (short) length, can be identified. A phase space point which belongs to an unsuccessful transition trajectory, which is short in time, in a system with a lowered barrier is less likely to be on a successful transition path of the same length in the original system with a higher barrier. As a result, the phase space points which are more likely to be "reactive" will be chosen with a high probability for the forward/backward trajectory shooting. This procedure will further reduce the computational cost by reducing the sampling over the unsuccessful transition paths (the ratio between the successful and unsuccessful trajectories are easily recovered, see below).

The method described above is illustrated in Fig. 1 using a double well potential. The two states corresponding to the two potential wells are labeled as states A and B (the reaction coordinate is labeled as R.C.). First, a simulation is performed for the biased potential and successful trajectories such as the one shown at Fig.1(b) are obtained. Many small segments of the trajectory which contain the transition path (such as the one enclosed in the rectangular region at Fig. 1(b)) will be chosen for further simulations. Phase space points (configuration and momentum) will be randomly (or every certain number of points) taken from these segments, e.g., the points 1, 2, and 3 in Fig. 1(b). In the next step, MD simulations will be performed on the original potential, using each of the chosen snapshot as the starting point and the system will be propagated both forward and backward for a given short time period. If the resulted trajectory ends at potential well A (e.g., the backward trajectory) and B (e.g., the forward trajectory), it is counted successful

(e.g., trajectory 2 in Fig. 1(d)). Otherwise, it is considered unsuccessful (the same as in the standard TPS approach). With a sufficient sampling of these trajectories and phase space points, the successful transition paths can also be adequately sampled. The number of corresponding unsuccessful trajectories, N_{non}, that is needed for the rate calculations can be estimated as $N_s = N_{non}\eta$, where η is the ratio between N_s, the re-weighted number of phase points that leads to successful transition paths in the given propagating time and that of the unsuccessful ones,

$$\eta = \frac{\sum_{succ} e^{\beta f(V)}}{\sum_{unsuc} e^{\beta f(V)}} \tag{2}$$

In Eq. (2) the summation is over the initial shooting points corresponding to the successful and unsuccessful transition paths respectively. As usual, $\beta = k_B T$ and k_B is the Boltzmann constant and T is the temperature. This weighting factor, $\exp[f(V(r))/k_B T]$, where $f(V(r))$ is the bias potential, recovers the Boltzmann distribution of the original system. The rate constant is then calculated in the standard way of transition path sampling based on the numbers of successful and unsuccessful trajectories

$$k = \left(\frac{N_s}{N_{non} + N_s}\right)/\tau = \left(\frac{\eta}{1+\eta}\right)/\tau \tag{3}$$

where τ is the length of the trajectories. Equation (3) is obtained by assuming a first-order reaction and when the length of the trajectories is much shorter than the reaction half time.

Two aspects of the simulation should be checked for convergence. Firstly, the convergence of the potential of surface (or the distribution function of energy) is checked

for the efficient sampling of the configurations. In the implementation of this method, accelerated MD simulations will be performed and as a result, a reaction coordinate can be identified. Free energy of transition along the reaction coordinate is then computed till convergence (to guarantee that a correct sampling over the configuration space has been obtained). Secondly, the ensemble of initial configurations selected for transition path shooting will be varied (e.g., by changing the range of a loosely defined reaction coordinate) to test the convergence in rate constant calculation. We note here that the reaction coordinate is only used for the check of convergence of the configuration space sampling but it is not needed for the transition path sampling.

3. Application to Conformational Changes of Methyl Maltoside

In the following, as an example, we apply the selective sampling method of transition paths introduced above to study the conformational change of methyl maltoside (see Fig. 2), and compare our results with those from studies using the standard transition path sampling method.[42] All simulations were carried out using the AMBER8[43] ff94 all atom force field[44] in conjunction with the GLYCAM 04 parameters for carbohydrates.[45] As recommended, the AMBER 1-4 electrostatic and nonbond scale factors, scee and scnb, were set to 1.0 in order to properly treat internal hydrogen bonds. The generalized Born implicit solvent model[46,47] was used to model the effects of solvation. The SHAKE algorithm[48] with a relative geometric tolerance of 10^{-5} was used to constrain all bonds including hydrogen. The system was maintained at the objective temperature using the weak-coupling algorithm[49] with a coupling constant of 1.0 ps^{-1}. Firstly, short simulations were run at the desired temperature 298 K to yield information on energy distributions. Based on the knowledge of energy distributions, a bias potential in the form of Eq. 1

($n = 4,\ \ a_1 = -0.9,\ \ V_1 = 222,\ \ \sigma_1 = 23;\ \ a_2 = -0.24,\ \ V_2 = 241,\ \ \sigma_2 = 22;\ \ a_3 = -2.9,$ $V_3 = 251, \sigma_3 = 95; a_4 = -7.6, V_4 = 262, \sigma_3 = 65.$ a, V, σ are all in units kcal/mol) was added to broaden the energy distribution so that the high energies which are only significantly populated at high temperatures will also be efficiently sampled in a simulation at the lower temperature. Then six independent accelerated MD simulations using the bias potential were conducted and each one of them was extended to 400 ns. Since it enhances sampling for higher energies, the bias potential leads to more frequent transitions between different conformations. As an example, the motions along the dihedral angle ψ during a 20 ns segment of an accelerated MD simulation are shown in Fig. 3. Seven obvious transitions can be easily located in the figure. While in the normal MD simulation the transition along angle ψ is a rare event. Typically, a normal MD simulation would have to be run for 160 ns on average in order to see just one transition.[42] As mentioned earlier, the convergence of the potential of surface should be checked to guarantee that a correct sampling over the configuration space has been obtained. The free energy of transition along the dihedral angle ψ was calculated and shown in Fig. 4. The solid line represents the free energy calculated from accelerated MD simulations and the dashed line represents the free energy re-weighted by the weighting factor $\exp[f(V(r))/k_B T]$, which is the free energy should be experienced in normal MD simulations. Apparently, the free energy barrier between conformation B and F (as defined in Ref. 42) has been effectively lowered so that the transition occurred more frequently in accelerated MD simulations. After being re-weighted from the result of accelerated MD simulations, a free energy barrier of 6.9 kcal/mol was obtained between B and F from a negative rotation about ψ angle from B. And the same transition in the

positive ψ direction has a free energy barrier 10.9 kcal/mol. These results were in good agreement with previous umbrella sampling[42] and accelerated MD simulation results.[27]

In order to combine accelerated MD simulations and transition path shooting method, potential energies, coordinates and velocities were recorded every 50 steps in all six simulations (every one out of 50 steps is used to reduce the correlation of the selected points). In practice, choosing fewer points will further reduce the correlation between different trajectories but also reduces the number of successful trajectories. In the simulation, there were total 4.8×10^7 conformations which consisted of the ensemble of starting points for shooting trajectories using the selective sampling method of transition paths. Certainly, there is no need to shoot trajectories from all these points. The benefit of using selective sampling of transition path method is that the most "reactive" phase space region can be identified through the biased sampling of the phase space. Focusing on these points in the identified "reactive" phase space portion will reduce the computational cost on the sampling over the unsuccessful transition paths. In our simulations, an average of 1715 points near reactive phase space region was selected out of 4.0×10^5 phase space points obtained from a 20 ns accelerated MD simulation. And 205792 points in total (0.43%) were focused on out of the ensemble consisted of 4.8×10^7 phase space points. After shooting backward/forward from these 205792 points, 10729 successful transition trajectories were obtained. The total simulation time is then approximately 3.2 μs (the sum of the time needed by accelerated MD simulations: $6\times400\,ns = 2.4\,\mu s$ and trajectory shooting time: $205792\times4\ ps \sim 823\,ns$). While, during the same simulation time, only ~20 transitions could be found in normal MD simulations. Although not all of the 10729 transition paths are important at room temperature, the improvement of the

efficiency in sampling transition paths is prominent.

In the selective trajectory sampling, sets of coordinates and velocities corresponding to these selected points were combined to form the restart files for trajectory shooting forward and backward (reverse the direction of stored velocities) with system energy being kept constant. New paths were generated under the original potential, using each of the chosen snapshots as the starting point and allowing it to propagate both forward and backward in time for 2 *ps*. The next step was to determine whether this new transition path was successful or not. According to the free energy calculation (Fig. 4), the reactant region was defined as $-70^\circ < \psi < 40^\circ$ and the product region was defined as $\psi > 150^\circ$ or $\psi < -150^\circ$. If the backward segment of a trajectory ends at the reactant region and the forward segment of the same trajectory ends at the product region, it is counted successful. Otherwise, it is considered unsuccessful. Fig. 5 shows two typical successful transition paths. In Fig. 5, the starting points are in the middle (at time *t=0*). Trajectory segments of shooting backward are represented by grey solid lines and segments of shooting forward are shown with black solid lines. Connecting a pair of grey and black lines shows a 4 *ps* transition path. In Fig. 5(a), starting from point at $\psi = -107^\circ$, the forward segment ends at the product region F in 2 *ps* and the backward segment ends at the reactant region B in 2 *ps*. This is a successful 4 *ps* transition path passing through transition state T_1 (defined in Fig. 4) by a rotation about the ψ dihedral angle in a negative direction (ψ decreases). In Fig. 5(b), the starting point is at $\psi = -100.6^\circ$. Connecting backward and forward segment shows a successful 4 ps transition path passing through transition state T_2 (Fig. 4) by a rotation about the ψ dihedral angle in a

positive direction. Fig. 6 shows two typical unsuccessful transition paths. In Fig. 6(a), although the backward segment ends at reactant region B, the forward segment does not reach the product region F yet. While in Fig. 6(b), the forward segment ends at the product region F but the backward segment does not reach the reactant region B. Therefore, they are both unsuccessful paths. In this work, three different ensembles of initial configurations were used for transition path shooting in order to check the convergence in rate constant calculation. The first initial configuration ensemble included points near reactive regions $90^\circ < \psi < 140^\circ$ and $-150^\circ < \psi < -90^\circ$ in successful transitions under biased potential. Firstly, points in these two regions were selected. Then forty points (equal to 2 *ps* in time) before and forty points after each point of these pre-selected points were picked. In the selection of the second initial configuration ensemble, the reactive region with negative ψ was enlarged a little bit: $-150^\circ < \psi < -80^\circ$. The other selecting criteria were kept the same as the first ensemble. And, in the third ensemble, the definition of reactive regions were even looser: $80^\circ < \psi < 140^\circ$ and $-150^\circ < \psi < -80^\circ$. Also, sixty points (3 *ps* in time) before and after the pre-selected points were included. For the first initial configuration ensemble, the rate constant of conformational change between B state and F state was estimated as $9.93(\pm 2.50)\times 10^6\ s^{-1}$. Apparently, this is not converged yet. The rate constant estimated from transition path shooting simulation using the second and the third initial ensemble was $1.22(\pm 0.39)\times 10^7\ s^{-1}$ and $1.26(\pm 0.31)\times 10^7\ s^{-1}$, respectively. Further expansion of the initial configuration ensemble (including more points) did not affect the results. So the rate constant, $1.26(\pm 0.31)\times 10^7\ s^{-1}$, is the estimation of the rate constant of conformational change between B state and F state which is in reasonable agreement with

the previous studies by transition state theory and transition path sampling method.[27,42] The probability density of states over φ and ψ was calculated from the successful transition path ensemble and shown in Fig. 7. The larger number and warmer color in Fig. 7 shows a greater probability. All three natural transition path ensembles found in the previous TPS study[42] were also captured in this study. The most favorable transition path is the transition path E1 observed in Ref. 42. In this transition path ensemble transitions propagated from state B to state F through a rotation about the ψ dihedral angle in a negative direction and passed through the transition state T_1 (see Fig. 5(a)). A less favorable transition path ensemble is E2. This class of transition paths also propagated from state B to state F through a rotation about the ψ dihedral angle. But it passed through a metastable region[42] G $(+30^\circ,-130^\circ)$ instead of T_1. The third transition path ensemble is the transition path E4 in which transitions propagated from state B to state F through a rotation about the ψ dihedral angle in a positive direction and passed through the transition state T_2 (see Fig. 5(b)). Since the free energy barrier between B and F through a rotation in a positive direction is higher than that through negative direction (Fig. 4), transition path E4 has the lowest probability density in Fig. 8. This means that it is the most unfavorable transition path ensemble. The advantage of selective sampling of transition path method is that it is relatively easy to identify different reaction channels and obtain their statistical correlation, since the sampling of the most reactive phase space region is efficient. Although theoretically TPS is also able to sample all of phase space efficiently, it is practically very difficult. For example, transition path ensemble E1, E2, E4 were all identified in the earlier TPS study, but TPS will need additional assistance, such as an alternative initial path making the transition from B to F through a rotation of ψ in a positive direction, in order to sample the transition path E4.[42] Unfortunately, the

statistical correlation between the two different paths is then lost in this case.

4. Discussion and Conclusion

In this short paper, we present a new and simple trajectory shooting method for the sampling of transition paths and for the calculation of rate constants. The method makes use of the accelerated MD simulations to enhance the sampling of the "reactive" configuration space through adding a bias potential to the original potential. The accelerated MD simulation is especially helpful for systems with high energy barriers, in generating the initial phase space points for a selective sampling of successful transition paths. The effectiveness of the method depends on the selection of the bias potential. An easy way of selecting bias potential based on the potential energy information obtained from short normal MD simulations has been provided. The method accelerates the transition path sampling and rate constant calculations by significantly reducing the sampling of non-reactive trajectories, and at the same time calculates the ratio between reactive and non-reactive transition paths quickly. Further, as a result of these simulations, both thermodynamic and dynamical information can be easily obtained for the original system at all desired temperatures (as long as the distribution of sampled energy covers the energy distribution at the given temperature). The present method also conserves in part the "throwing rope in the dark" advantage of TPS in that it does not require the knowledge on the reaction coordinate and thus is potentially useful for systems without known reaction coordinates and for those with multiple reaction paths. The method is expected to be useful in identifying new reaction pathways. The main difference between the present method and the standard TPS is that instead of generating transition paths from a pre-selected reactive one, the shooting of trajectories was initiated

from an ensemble of phase space points. The sampling over reactive transition paths is enhanced through the enhanced sampling of initial configurations that more likely lead to reactive paths. The enhanced sampling of the more reactive region reduces the computational requirement of the large less reactive phase space and provides a quick estimate of the rate constant, although it might slow down the convergence in the rate constant calculation, due to the neglect of these less reactive points (not the rather large error bar of the calculated rate constant). Since a large sample of uncorrelated trajectories is needed in this method, an analysis of the minimum size of the sample generated using the biased potential is an important issue. However, an exact theoretical analysis is difficult. Only a rough estimation can be given as guidance. In practice, the convergence of the calculated results has been tested numerically based on simulations using different trajectories and on the length of trajectories. Finally, we note that the enhanced sampling is performed in a controlled fashion so that the Boltzmann distribution and the unbiased sampling of transition paths are easily recovered for the real system.

We also note here that alternative approaches may be used in the combination of accelerated MD simulations with the transition path sampling. For example, one can choose to relax the trajectories obtained from the accelerated MD simulation and then perform the standard TPS, similar to the approach used by Hu *et. al*. in steered molecular dynamics simulations.[38] Parallel TPS can also be performed using a number of different initial biased trajectories, which are expected to speed up the simulations and avoid the trapping of transition paths through a local transition state when multiple transition states/trajectory ensembles exist. The method is expected to have two advantages over the standard transition path sampling method: (1) it provides a quick method for generating

the initial trajectories. (2) Since several initial trajectories are generated from the biased potential and are uncorrelated, the sampling efficiency is expected to be further enhanced.

Acknowledgement

YQG acknowledges ACS-PRF and the Camille and Henry Dreyfus Foundation for support. YQG is a 2006 Searle Scholar.

References:

[1] D. D. Frantz, D. L. Freeman, and J. D. Doll, J. Chem. Phys. **93**, 2769 (1990).

[2] C. Bartels and M. Karplus, J. Phys. Chem. B **102**, 865 (1998).

[3] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).

[4] A. Mitsutake and Y. Okamoto, Chem. Phys. Lett. **332**, 131 (2000).

[5] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).

[6] M. Falcioni and M. W. Deem, J. Chem. Phys. **110**, 1754 (1999).

[7] Q. L. Yan and J. J. de Pablo, J. Chem. Phys. **111**, 9509 (1999).

[8] Q. L. Yan and J. J. de Pablo, J. Chem. Phys. **113**, 1276 (2000).

[9] G. Stolovitzky and B. J. Berne, Proc. Natl. Acad. Sci. U. S. A. **97**, 11164 (2000).

[10] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).

[11] Y. Sugita, A. Kitao, and Y. Okamoto, J. Chem. Phys. **113**, 6042 (2000).

[12] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991).

[13] S. G. Itoh and Y. Okamoto, J. Chem. Phys. **124**, 104103 (2006).

[14] G. Bussi, A. Laio, and M. Parrinello, Phys. Rev. Lett. **96**, 090601 (2006).

[15] H. Grubmuller, Phys. Rev. E **52**, 2893 (1995).

[16] J. Lee, H. A. Scheraga, and S. Rackovsky, J. Comput. Chem. **18**, 1222 (1997).

[17] A. F. Voter, Phys. Rev. Lett. **78**, 3908 (1997).

[18] A. F. Voter, J. Chem. Phys. **106**, 4665 (1997).

[19] L. Piela, J. Kostrowicki, and H. A. Scheraga, J. Phys. Chem. **93**, 3339 (1989).

[20] C. H. Mak, J. Chem. Phys. **122** (2005).

[21] L. D. Gelb, J. Chem. Phys. **118**, 7747 (2003).

[22] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).

[23] D. Hamelberg, J. Mongan, and J. A. McCammon, J. Chem. Phys. **120**, 11919 (2004).

[24] D. Hamelberg, T. Shen, and J. A. McCammon, J. Chem. Phys. **122**, 241103 (2005).

[25] D. Hamelberg, T. Shen, and J. A. McCammon, J. Am. Chem. Soc. **127**, 1969 (2005).

[26] D. Hamelberg and J. A. McCammon, J. Am. Chem. Soc. **127**, 13778 (2005).

[27] Y. Q. Gao and L. J. Yang, J. Chem. Phys. **125**, 114103 (2006).

[28] L. J. Yang, M. P. Grubb, and Y. Q. Gao, J. Chem. Phys. **126** (2007).

[29] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).

[30] P. G. Bolhuis, C. Dellago, and D. Chandler, Faraday Discuss., 421 (1998).

[31] C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **110**, 6617 (1999).

[32] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[33] P. G. Bolhuis, Proc. Natl. Acad. Sci. U. S. A. **100**, 12129 (2003).

[34] J. Marti and F. S. Csajka, Physical Review E **69** (2004).

[35] T. S. van Erp, D. Moroni, and P. G. Bolhuis, J. Chem. Phys. **118**, 7762 (2003).

[36] D. Moroni, T. S. van Erp, and P. G. Bolhuis, Physica A **340**, 395 (2004).

[37] D. Moroni, P. G. Bolhuis, and T. S. van Erp, J. Chem. Phys. **120**, 4055 (2004).

[38] J. Hu, A. Ma, and A. R. Dinner, J. Chem. Phys. **125** (2006).

[39] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U. S. A. **97**, 5877 (2000).

[40] R. Radhakrishnan and T. Schlick, Proc. Natl. Acad. Sci. U. S. A. **101**, 5970 (2004).

[41] Constant energy shooting is used mainly because: I. During the short time shooting, the change of energies along the transition paths is small since the energy relaxation in this short time is minimal. We have also tried constant temperature shooting and the results are similar, especially when a weak coupling method, as done in the current study, is used to keep the constant system temperature. II. Since the system size is limited, and approximations have been used for the coupling between the environment and the system to keep the constant temperature, both constant temperature and constant energy calculations only give approximately correct transition paths (the real coupling between the system and the surroundings is not taken into account precisely). The constant energy shooting is sufficient to provide a quick estimate of the rate constant.

[42] R. J. Dimelow, R. A. Bryce, A. J. Masters, I. H. Hillier, and N. A. Burton, J. Chem. Phys. **124** (2006).

[43] D. A. Case, *AMBER 8*. (University of California, San Francisco., 2004).

[44] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).

[45] R. J. Woods, R. A. Dwek, C. J. Edge, and B. Fraserreid, J. Phys. Chem. **99**, 3832

(1995).

[46] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, J. Am. Chem. Soc. **112**, 6127 (1990).

[47] V. Tsui and D. A. Case, Biopolymers **56**, 275 (2000).

[48] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).

[49] H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).

Figure Captions:

Figure 1. Scheme of the selective sampling of transition path method: (a) A double well potential model system. (b) The barrier in the original potential (black solid line) is lowered in the biased potential (black broken line). A successful transition trajectory under bias potential is shown with the gray arrow line. Points such as 1, 2, 3 in the dotted rectangular region were chosen for trajectory shooting. (c), (d), (e) Shooting trajectories starting from points 1, 2, 3 in (b). Gray arrow lines represent the backward shooting trajectories and black arrow lines denote the forward shooting trajectories.

Figure 2. Structure of methyl maltoside.

Figure 3. Motions along the dihedral angle ψ during a 20 ns segment of an accelerated MD simulation.

Figure 4. The potential of mean force of maltose molecule as a function of ψ: the solid line (accelerated MD) and the broken line (re-weighted from the accelerated MD by a weighting factor $\exp[f(V(r))/k_B T]$.

Figure 5. Typical successful transition trajectories: (a) successful transition through the transition state T_1 (see text), (b) successful transition through the transition state T_2 (see text).

Figure 6. Typical unsuccessful trajectories: (a) forward segment does not reach product region, (b) backward segment does not reach reactant region.

Figure 7. Probability density of states over φ and ψ calculated from successful transition path ensembles. Three transition path ensembles are labeled as E1, E2, E4.

Figures:

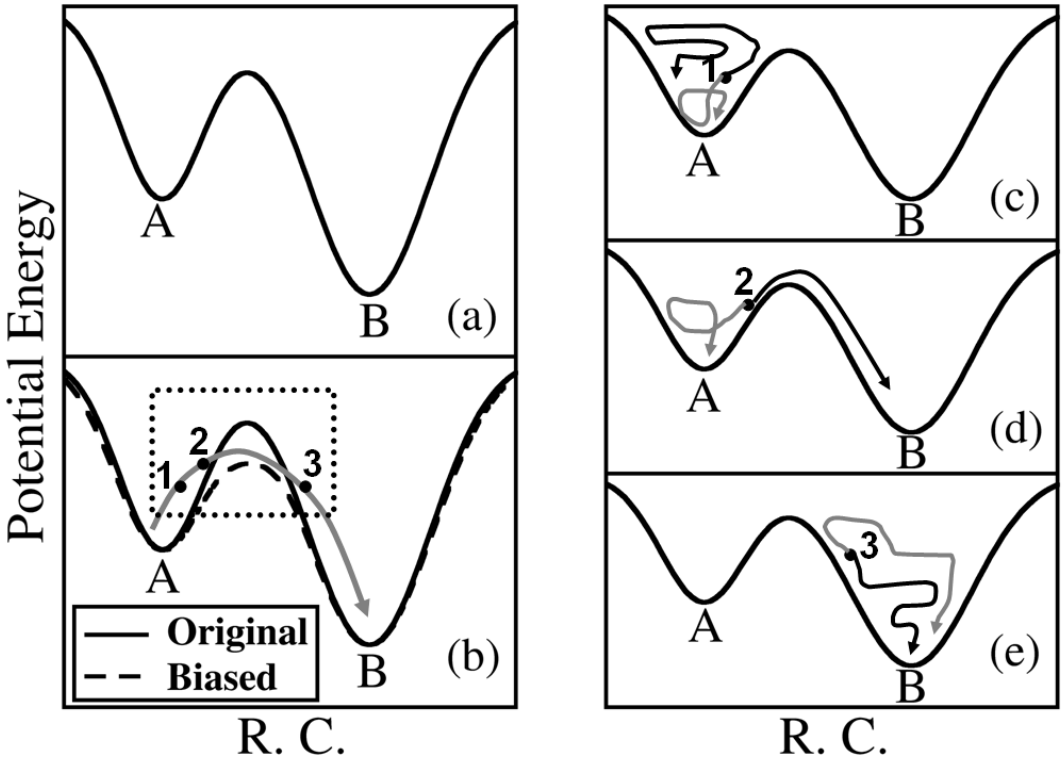

Figure 1

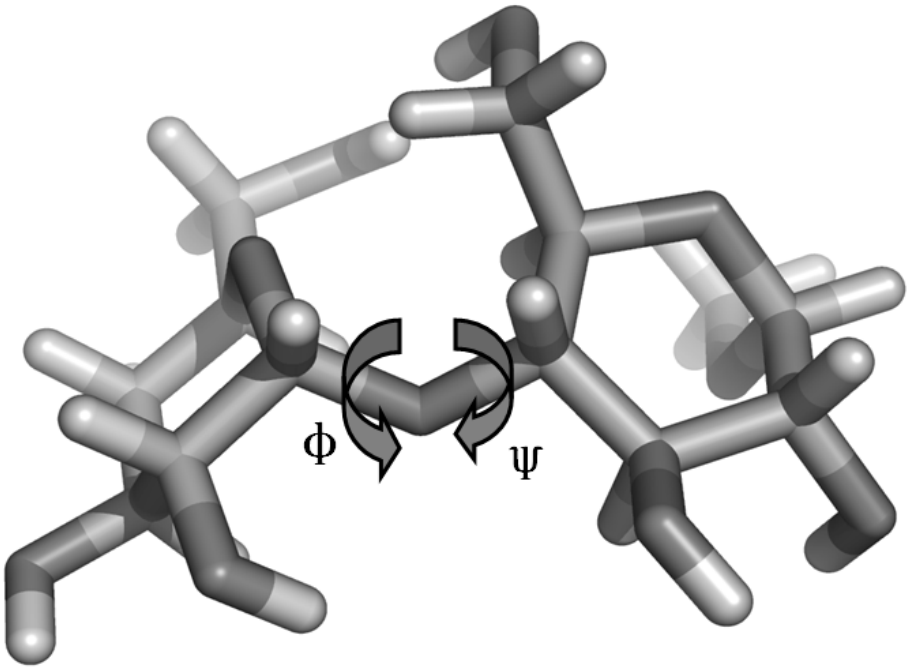

Figure 2

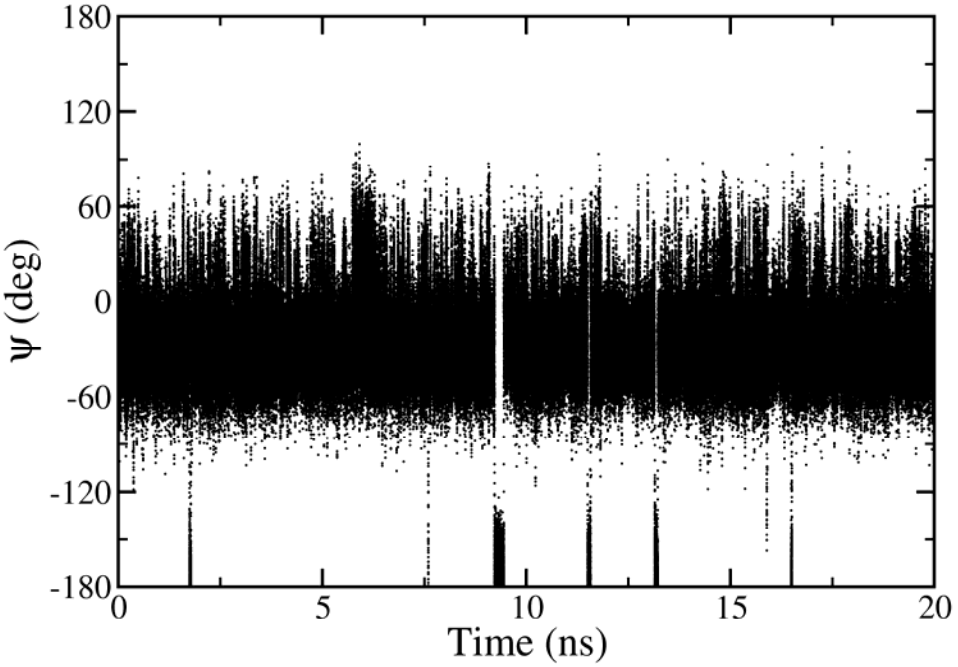

Figure 3

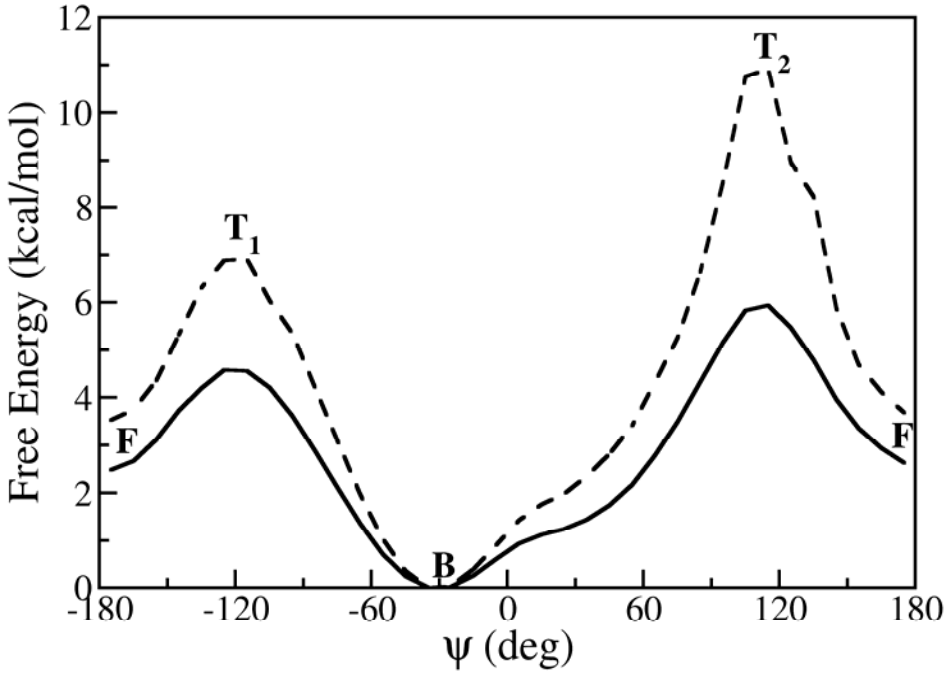

Figure 4

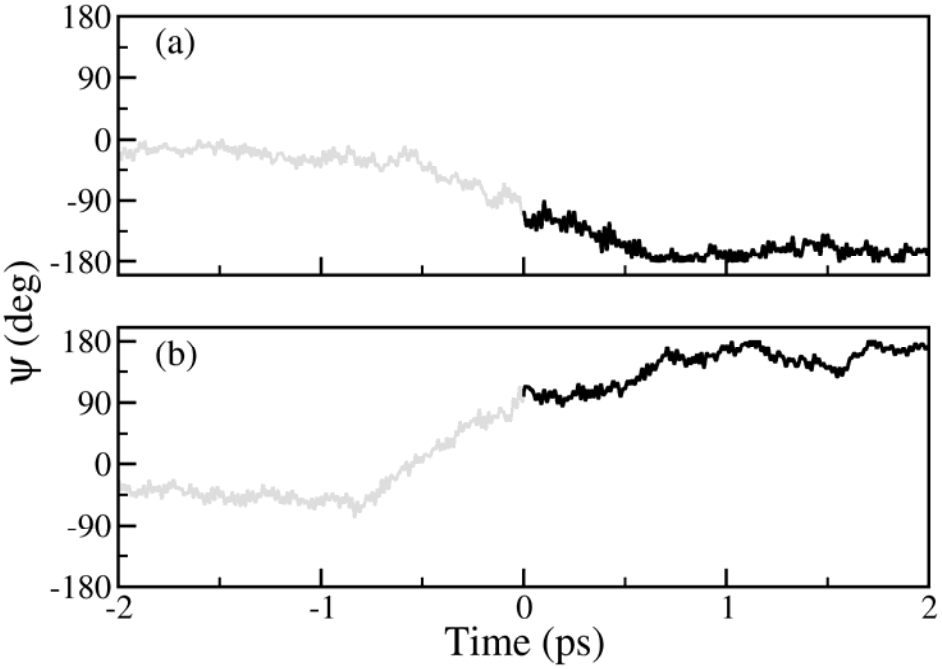

Figure 5

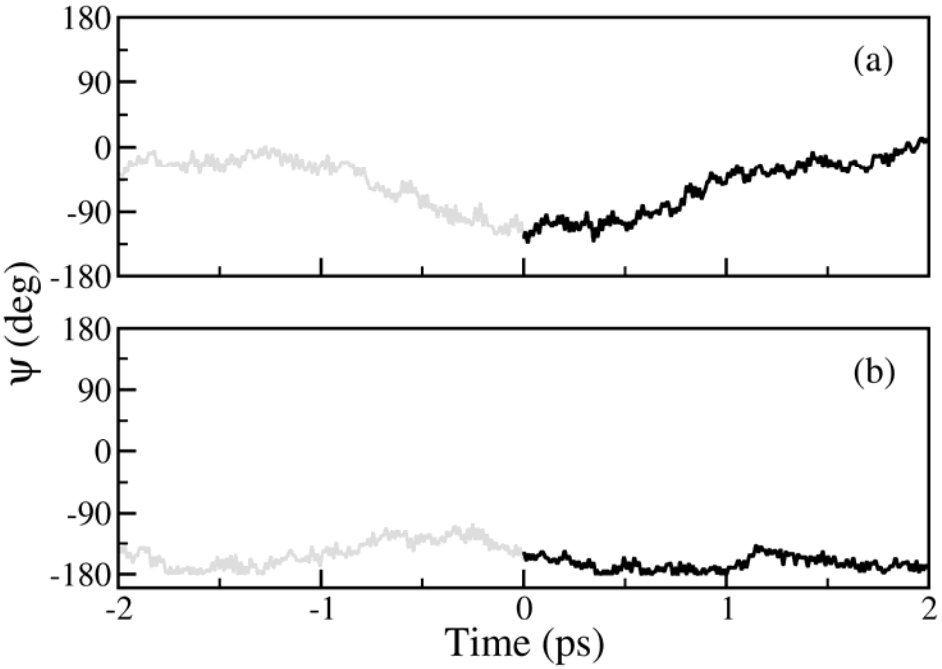

Figure 6

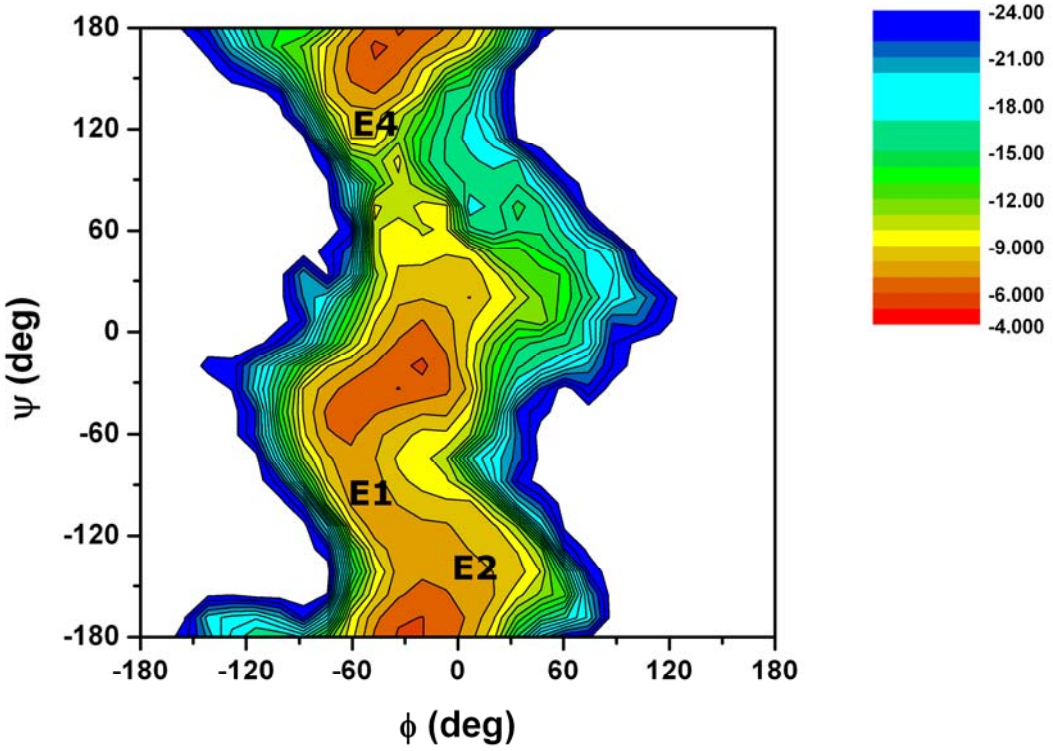

Figure 7

VITA

Name:	Xuebing Fu
Address:	Department of Chemistry 1124 Chemistry Bldg. Texas A&M University College Station, TX 77842
Email Address:	xuebingfu@yahoo.com.cn
Education:	B.S., Chemistry, Shandong University, China, 2006